Supplementary Methods

1.1. *Samples*

We used the publicly available aligned reads for the Ashkenazim trio (NA24385/HG002, NA24149/HG003, NA24143/HG004) from the Genome In A Bottle (GIAB) Consortium. We selected the 300x Illumina HiSeq paired end sequencing data downsampled to 60x and the 10X Genomics Chromium dataset both for improved mapping using linked reads and the ability to separate the reads by haplotype. We used the pre-aligned bam files available from `ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/`.

1.2. *Determination of Putative CNV sites*

To form a preliminary integrated deletion callset for the Ashkenazim trio son (NA24385/HG002/RM 8391), we used callsets submitted to Genome In A Bottle (GIAB) from multiple technologies as of Aug 2016 for GRCh37. The calls are available at `ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/analysis/NIST_DraftIntegratedDeletionsgt19bp_v0.1.7/`. GIAB used a series of heuristics to integrate deletions from these callsets:

(1) Find regions with deletion calls within 50bps of each other (i.e. sites are merged if they overlap by any amount or have <50bp separating them using bedtools merge -d 50).
(2) Annotate each region with the fraction of bases covered by calls from each method.
(3) Divide call regions into those that contain >25% tandem repeats (TR bed file with regions >200 bps in length containing TRs with >95% identity: `https://github.com/ga4gh/benchmarking-tools/blob/master/resources/stratification-bed-files/LowComplexity/AllRepeats_gt200bp_gt95identity_merged.bed.gz`). Remove all calls with TRs.
(4) Find regions for which methods from at least 2 technologies have <20% difference in predicted size, treating all methods equally. The methods considered from each technology are as follows:
    - Illumina: spiral, cortex, commonlaw, MetaSV, Parliament/assembly, Parliament/assembly-force, CNVnator, GATK-HC, freebayes
    - PacBio: CSHL-assembly, sniffles, PBHoney-spots, PBHoney-tails, Parliament/pacbio, Parliament/pacbio-force, MultibreakSV, smrt-sv.dip, Assemblytics-Falcon, Assemblytics-MHAP
    - Complete Genomics: CG-SV, CG-CNV, CG-vcfBeta
    - Optical Mapping: BioNano, BioNanoHaplo (haplotype-aware)
(5) Find the sensitivity of each method to these calls in the size bins:
    - 20-49bp
    - 50-100bp
    - 100-1000bp
    - 1000-3000bp

- >3000bp

(6) Filter all calls for which there is an overlapping call from any method that differs in size by >2x. Only include callsets if 75% of calls are <20% different from median size in the <3kb or >3kb size range (whichever the call falls in). This is intended to remove calls that may be more complex than just a simple deletion (e.g. a compound heterozygous event with deletions of different lengths or deletions of ambiguous length in repetitive regions).
(7) Filter calls overlapping segmental duplications >10kb by 20% or more, or overlapping Ns in GRCh37 by any amount.
(8) Use the breakpoints from the caller with the best size accuracy as described below.

Process for finding accuracy of predicted size from different techniques:

(1) Find regions with support from >3 callers.
(2) Find the difference between the size predicted by each caller and the median size of all callers with calls in each region.
(3) Find the 75th percentile of this difference (i.e. 75% of calls have a predicted size $< x\%$ different from the median size).

Note that this is likely an overestimate of the size accuracy due to the limited number of callsets used and it does not account for any biases that might be present in all or most callsets.

1.3. *Selection of CNV sites to show*

For the pilot study we randomly sampled 500 CNV sites to use in our experiment to represent a diversity of putative site sizes and support. We binned all putative CNV sites into 5 size groups: 100-299bp, 300-399bp, 400-999bp, 1000-1999bp, 2000-2999bp. We sampled 100 sites from each size bin and within these size bins sampled 60 sites supported by 3 or more technologies, 20 supported by 2 technologies and 20 supported by 1 technology. We did no additional filtering of these randomly sampled sites. For the genome-wide set we selected all putative CNV sites supported by at least 2 technologies and within 100-2999bp that were not profiled in the pilot study.

1.4. *Experimental Design*

We represented each site with six different images in the CrowdVariant platform. We used 10X data for three images and Illumina paired-end data subsampled at 60x for the other three. The three images displaying 10X data showed the son (NA24385/HG002) as well as the same reads split by haplotype, each shown in a separate image. The three images using Illumina data displayed the family trio of son, father and mother (NA24385/HG002, NA24149/HG003, NA24143/HG004) at the same site. While PacBio reads are also known to highlight SVs well, we found through test documentation that non-experts found the visualization substantially more confusing.

We asked for 20 different non-expert workers to classify each image and 5 different experts to classify the same images. For the experts, we batched the 500 sites into 5 sets of 100 and prioritized classifying all 6 images for the first set of 100 sites before moving to the next 100 sites. Each worker could only classify the same site once, and the images were presented in a random order to each worker within each batch.

1.5. *Image Generation*

We used IGV to generate images for each site. We showed the site as well as 80% of the size of the region on either side for genomic context.

1.6. *Participant Recruitment*

Non-experts were recruited from Google's CrowdCompute team. This group of workers is untrained in genomics applications, but has experience in other crowdsourcing tasks. We recruited experts from the genomics community via email and social media. Workers answered as many questions as they were willing. Experts were determined by self-reporting some or substantial experience classifying CNVs.

1.7. *Participant Instruction*

Workers were shown an example pileup image highlighting three features useful to focus on: the putative CNV site location, the aligned reads and the overall distribution of reads over the reference genome. We showed two examples images each of copy number 0, copy number 1, and copy number 2. We also showed two example images each of accurate and inaccurate putative sites. The same training images were shown to both experts and non-experts, but the non-experts were shown modified language that did not contain technical terms (e.g. bars instead of reads). We did not bring additional attention to IGV features that may help image classification so that the documentation was the same for both groups, but experts may have made use of additional visual cues in the image given prior knowledge such as coloring for abnormally spaced reads or variant annotations.

1.8. *Crowdsourcing Platform*

We used Google's Crowd Compute crowdsourcing platform. We developed a plugin that shows one image at a time along with a set of three questions. The image was not labeled with coordinates or any other identifying information beyond what is shown in IGV. Each worker answered 3 multiple choice questions about each image. We listed the true copy number next to the description of the pattern (Missing/Half/Complete) to use the same plugin for both experts and non-experts. For haploid images, the patterns should only resemble Missing or Complete although technically Complete (copy number 2) is not correct. In future iterations we would recommend focusing on the visual pattern or the true copy number status in the answer choices to avoid confusion.

(1) What is the status of the genome under the blue bar?

Missing (copy number 0)
Half (copy number 1)
Complete (copy number 2)
None of the above

(2) How confident are you?

Very unsure
Slightly unsure
Somewhat sure
Very sure
Extremely sure

(3) Is the blue bar accurate?

Accurate
Inaccurate
Not applicable

The answer for each question as well as the time to answer the questions was recorded. Workers were allowed to skip questions. We collected answers from the external community over 10 days.

1.9. *Baseline voting model*

We scored each possible copy number classification (CN0/CN1/CN2/None of the Above) at each site to determine the likely true copy number state as well as an associated level of confidence. We pursued several methods from a baseline voting scheme to methods that assess and re-weight some combination of each worker's ability, the usefulness of each image type (i.e. sequencing platform) and the likelihood of mistaking one type of CNV for another. While the weighting framework appropriately minimizes the influence of ineffective workers and more difficult image types, we ultimately found that its benefits were minimal compared to a simple baseline voting scheme and we moved forward with a voting model for simplicity. In situations with larger numbers of more variable workers, the benefits of a more complex modeling scheme would likely be greater.

The baseline voting scheme scores each copy number state as the percentage of worker responses that voted in favor of that classification. For the diploid images, we included all worker classifications. However, the haplotype image classifications were only counted if the classifications for the paired haplotype images were genetically plausible (no Mendelian violation). We initially allowed haploid CN0 and CN0 to count for diploid CN0, haploid CN0 and CN2 to count for diploid CN1 and haploid CN2 and CN2 to count for diploid CN2. Only haploid "None of the Above" and "None of the Above" counted toward "None of the Above". All other configurations (e.g. haploid CN0 and "None of the Above") were omitted from the vote. However, we observed a small boost in performance by allowing non-expert CN1 to count for CN2 in the haploid data since these errors were made systematically. Thus in the final model we also allowed haploid CN0 and CN1 to count for diploid CN1 and haploid CN1 and CN1 to count for diploid CN2. Using this voting model, we calculated for each site and for each individual in the trio a score for CN0, CN1, CN2 and "None of the Above" that sum to 1. The CrowdVariant score is the max of these scores.

1.10. *Weighted Framework*

We designed a model with three parameter sets based on our observations of the data. We note that despite using the same documentation, some non-experts appear to be more accurate than others. In addition, some of the images - whether due to sequencing depth, alignment quality or other factors - make it easier to determine the true copy number state from a pileup image. Finally, upon observing that workers can systematically mistake one type of copy number event for another, we note that we should be able to compensate for these types of systematic errors. Given these considerations, we define the score of each site to be a function of all the classifications for that site weighted by the worker ability and image usefulness and subject to systematic errors indicated by a confusion matrix mapping each classification to its likely intended classification.

Because we model latent diploid copy number state, we need to map the haplotype classifications to the diploid latent states that they support. We introduce the genetic function G in order to achieve that mapping. For diploid states, the genetic function returns the same state. For haploid data considered together, the genetic function will return the diploid state the data supports. Two haploid CN0s support diploid CN0 while haploid CN0 and CN2 support diploid CN1 and haploid CN2s support diploid CN2. When deciding how to translate haploid classifications with uncertainty (None of the Above answers) or discordant haplotypes (such as CN1 and CN1), we allowed two haploid "None of the Above" to contribute to "None of the Above" but otherwise prevent any discordant haplotype information from contributing to a classification.

With our three parameter sets, genetic function and observed and latent data, we describe the model updates below:

- Iterables: sites S, individuals I, genotypes G, workers W, image types M
- X: observed classifications for each site by each worker
  indexed X^w_{ism}
- C: latent true CNV state for each site
  indexed C^g_{is}
- θ: parameters
  W: worker ability
  indexed W_w
  A: image usefulness
  indexed A_m
  M: confusion matrices
  indexed $M^m_{g_X,g_I}$

$$P(C_{is} = g | X_{is}; \theta) = \frac{\sum_w^W W_w \sum_m^M A_m \sum_h^G M_m(X_{ism}^w, h) G(m, g, h) \mathbb{1}[X_{ism}^w]}{\sum_w^W W_w \sum_m^M A_m \sum_h^G M_m(X_{ism}^w, h) \mathbb{1}[X_{ism}^w]} \tag{1}$$

$$W_w = \frac{\sum_i^I \sum_s^S \sum_g^G P(C_{is} = g) \sum_m^M A_m \sum_h^G M_m(X_{ism}^w, h) G(m, g, h) \mathbb{1}[X_{ism}^w]}{\sum_i^I \sum_s^S \sum_g^G P(C_{is} = g) A_m \sum_h^G M_m(X_{ism}^w, h) \mathbb{1}[X_{ism}^w]} \tag{2}$$

$$A_m = \frac{\sum_i^I \sum_s^S \sum_g^G P(C_{is} = g) \sum_w^W W_w \sum_h^G M_m(X_{ism}^w, h) G(m, g, h) \mathbb{1}[X_{ism}^w]}{\sum_i^I \sum_s^S \sum_g^G P(C_{is} = g) \sum_w^W W_w \sum_h^G M_m(X_{ism}^w, h) \mathbb{1}[X_{ism}^w]} \tag{3}$$

$$M_m(g_X, g_I) = \frac{\sum_i^I \sum_s^S \sum_g^G P(C_{is} = g) \sum_w^W W_w A_m G(m, g, g_I) \mathbb{1}[X_{ism}^w = g_X] \mathbb{1}[X_{ism}^w]}{\sum_i^I \sum_s^S \sum_g^G P(C_{is} = g) \sum_w^W W_w A_m \mathbb{1}[X_{ism}^w = g_X] \mathbb{1}[X_{ism}^w]} \tag{4}$$

Noting that each parameter as well as our latent copy number state variables are defined in terms of one another, we use an iterative procedure to update each parameter and latent states one at a time. We initialize with equal worker abilities and equal image usefulness and confusion matrices as identity matrices.

We tried several iterations of the model updating all the parameters separately and in various combinations together. Upon updating all parameters, we discovered quickly that updating the confusion matrix based on highly variable classification data leads to severe over fitting of the data. We thus fixed the confusion matrices as the identity matrices for diploid classifications and to fix the confusion matrix for haploid sites allowing any CN1 to count toward a CN2. This means that haploid CN2/CN0 and haploid CN1/CN0 both count toward diploid CN1 and haploid CN2/CN2 and haploid CN1/CN2 both count toward diploid CN2.

1.11. *Metrics for evaluation*

There is no gold standard data available for all putative sites we evaluated. As a result, we rely on three metrics based on the familial structure of the 3 trio individuals to assess accuracy in addition to comparing to existing methods. First, we determined whether any Mendelian violations were made using the assigned most likely classifications for each site for each individual. For example, given a copy number 0 in the son, a copy number of 2 from either parent would be considered a Mendelian violation. For each site, we convert scores to hard classifications, determine whether a Mendelian violation was made and report the percent of sites that do not result in Mendelian violations. This metric has the caveat that some configurations, such as a CN1 in the son, are less likely to be caught if incorrect. However, the strong inverse correlation between our scores and Mendelian violations suggests it is sufficiently sensitive to evaluate our classifications as a whole.

In addition to the percentage of Mendelian violations, we assessed how well we could identify high confidence genetically plausible sites. We ranked the sites based on their CrowdVariant score (1 being most confident and 0.25 being least confident) and calculated the area under the ROC curve to assess how well our calculated scores discriminate genetically plausible from implausible sites. We also evaluated the uncertainty for the AUC using DeLong's method to estimate 95% confidence intervals (Robin et al., 2011) (DeLong, DeLong, & Clarke-Pearson, 1988).

For the third metric, instead of relying on the most likely copy number status for each site, we calculated the total probability of a Mendelian violation. For every violating configuration we added the product of the three probabilities that the members of the trio took that genetic configuration. This total genetic error allows more nuance in determining whether our improvements in estimated probabilities are more plausible even if the most likely site does not change. The total probability of a Mendelian violation was 14% for all 500 non-expert classified sites and 15% for the 100 expert classified sites [Supplementary Figure 6].

We compute each evaluation metric separately for the expert and non-expert data. Visualization of the ROC curve indicates that the small samples size of questions for the experts may make the AUC less reliable than for the non-experts, motivating our incorporation of 95% confidence intervals. For comparison, we randomized all answers by re-sampling the entire worker by classification matrices for both experts and non-experts and re-computed the same metrics. The rate of genetic plausibility was around 54% for randomized expert answers and 73% for randomized non-expert answers. The higher number of plausible sites for non-experts is likely due to the increased number of CN1 and smaller number of CN2 classifications they gave, which have more genetically plausible configurations.

1.12. *Comparison to svviz*

In order to assess evidence for each of the calls in this work and assign preliminary genotypes, we used svviz (Spies et al., 2015) to determine whether reads from five data sets support the reference allele, the alternate allele, or were ambiguous for each member of the trio. The five data sets used were (Zook et al., 2016):

- ∼300x 2x150bp Illumina paired end sequencing
- ∼45x 2x250bp Illumina paired end sequencing
- ∼10-15x 2x100bp Illumina mate-pair sequencing with ∼6kb insert size
- ∼25-60x 10X Genomics Chromium sequencing, with reads separated by haplotype using bamtools filter
- ∼30-70x PacBio Sequencing

We used the "ref_count" and "alt_count" outputs from svviz batch mode, which correspond to the numbers of reads unambiguously supporting the reference and alternate alleles, respectively. For the results of each dataset from all genomes, we visually examined the density of sites in a plot of log(alt_count) vs log(ref_count), as well as in histograms of alt_count and ref_count. Based on the density of sites in these plots, we chose cut-offs for alt_count and ref_count for each dataset to define likely homozygous reference (CN2), heterozygous (CN1), and homozygous variant (CN0) sites. If the site had few reads assigned to reference or alternate, or it it did not fall within the bulk of sites with any genotype, then it was assigned an ambiguous genotype for that dataset.

Dataset	HomRef minRef	HomRef maxAlt	Het minRef	Het minAlt	HomVar maxRef	HomVar minAlt
Illlumina 150bp	100	4	75	75	4	100
Illlumina 250bp	20	2	10	10	2	20
Illumina Mate-pair	10	40	20	20	10	40
10X	6*	1*	*	*	1*	6*
PacBio	10	0	5	5	0	10

A consensus genotype was then assigned if all datasets with an assigned genotype of CN0, CN1, or CN2 had the same genotype for that site. Sites with discordant genotype or with all ambiguous genotypes were given an uncertain consensus genotype. The uncertain classification in svviz is not directly comparable to "None of the Above" responses in CrowdVariant as some participants may have indicated a CN0/1/2 classification but marked low confidence to specify uncertainty while some may have specified "None of the Above."

When we computed the rate of Mendelian violations among the svviz classifications we found 489 out of 500 (97.8%) without Mendelian violations. One reason that svviz classifications have higher concordance is because there are larger number of "None of the Above" classifications, which cannot be used to detect Mendelian violations. In addition, we observe the same decreased rates of Mendelian violations when calculating site probabilities using only Illumina data (493 out of 500 (98.6%)). Based on careful examination of the data, we believe this is because images of the same type often show the same kinds of biases consistently among a family trio. A consistent bias from one platform can lead to increased concordance, but not necessarily an increase in ground truth accuracy. Combining high quality data from diverse sources should improve our best estimate of the truth. Further we are not using genetic concordance to capture ground truth for individual sites, but instead to determine a threshold at which we have high confidence in crowdsourced classifications and to compare experts and non-experts. When we remove all sites for which svviz reported an uncertain genotype for any individual, we observe a Mendelian concordance rate of 398/398 (100%). When we similarly remove any "None of the Above" classifications from CrowdVariant responses, we observe a Mendelian concordance rate of 423/458 (92.4%), but note that an overall larger number of sites are classified.

For more information, see
`ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/analysis/NIST_DraftIntegratedDeletionsgt19bp_v0.1.7/SVSummary_v0.1.7_TRs.html`.

1.13. *High confidence sites*

To determine high confidence sites we again ranked all of the 500 sites from the pilot study by the maximal score across all four classifications (CN0/CN1/CN2/None of the Above) for the son and identified the score at which the first Mendelian violation occurred. We determined all sites above this threshold as high confidence.

1.14. *Genome-wide analysis*

We selected all putative CNV sites supported by at least 2 technologies and not profiled in the first 500 sites (n=2271). We again recruited 20 classifications for all remaining sites by non-experts and used a baseline voting model to calculate final scores for each site. We used the threshold as determined in the pilot study to determine genome-wide high confidence sites. Using the same methods as described above for the pilot study, we evaluated the Mendelian violation rate, agreement with supporting technologies, and agreement with svviz.

Acknowledgments

We would like to thank Igor Karpov for his help working with the Crowd Compute infrastructure. We would like to thank Laura Paragano for her help in developing training materials. We are grateful to all participants who classified variants. We would like to thank Jonathan Bingham for his help in making this study possible. We would also like to thank the Verily analysis team for their feedback on the project. Certain commercial equipment, instruments, or materials are identified in this paper only to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the NIST, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

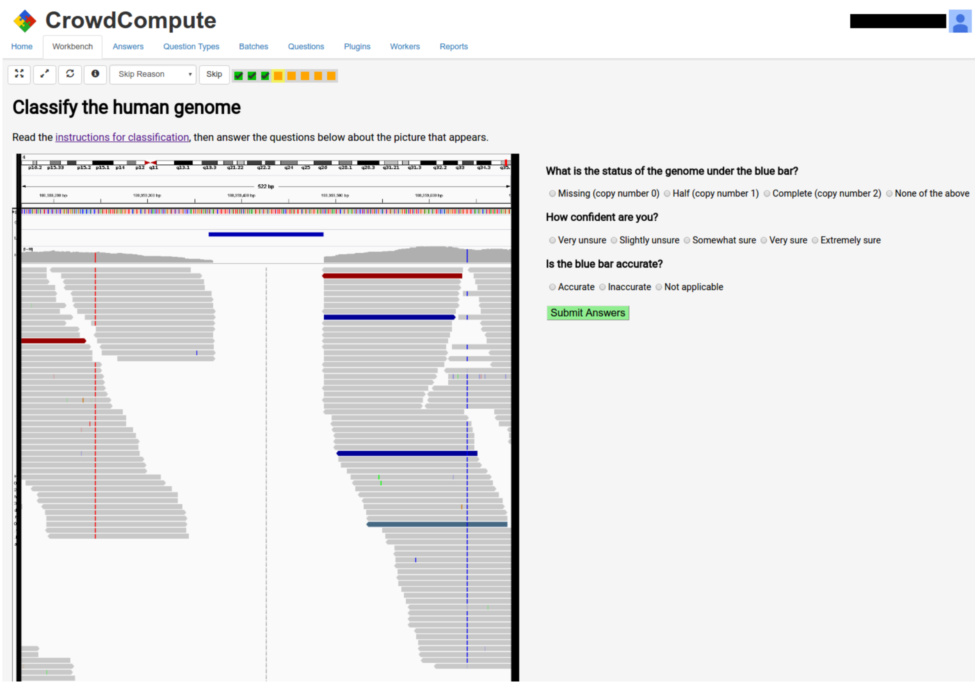
Author Contributions

P.G., J.Z., M.C., R.P., and M.D. designed the study. P.G. and M.D. developed the copy number variant crowdsourcing platform. J.Z. generated the results for orthogonal tools and determined putative CNV sites. P.G. developed the training and recruitment materials. P.G. performed analysis and modeling of the results. P.G., J.Z., M.C., M.S., and M.D. wrote the manuscript.

Disclosure Declaration

At the time of this work, M.D., R.P., M.C., and P.G. were employees of Verily Life Sciences, a for-profit corporation and affiliate of Google Inc. R.P. and M.D. are employees of Google Inc. Verily developed the CrowdVariant platform discussed in this paper, and could derive direct or indirect commercial benefit from positive research results pertaining to CrowdVariant.

Supplementary Figures



Supplementary Figure 1. A screenshot of the CrowdVariant crowdsourcing platform.

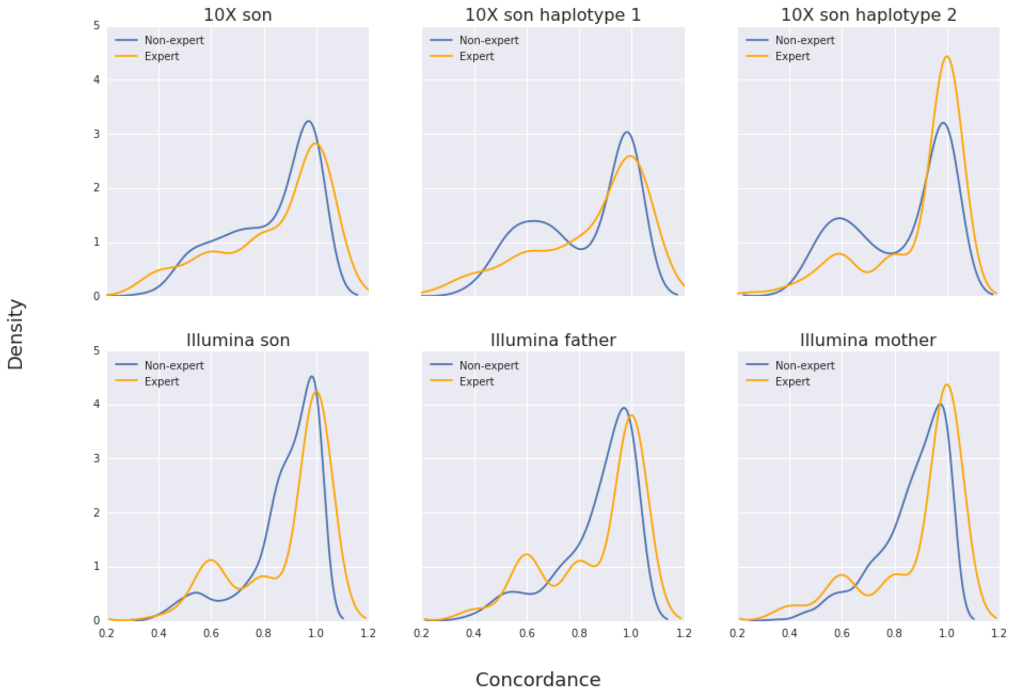
# Techs \ Size	100-299bp	300-399bp	400-999bp	1000-1999bp	2000-2999bp
3+	20	20	20	20	20
2	5	5	5	5	5
1	5	5	5	5	5

Supplementary Table 1. Number of sites sampled for each region size window and number of sequencing technologies supporting the putative CNV site.

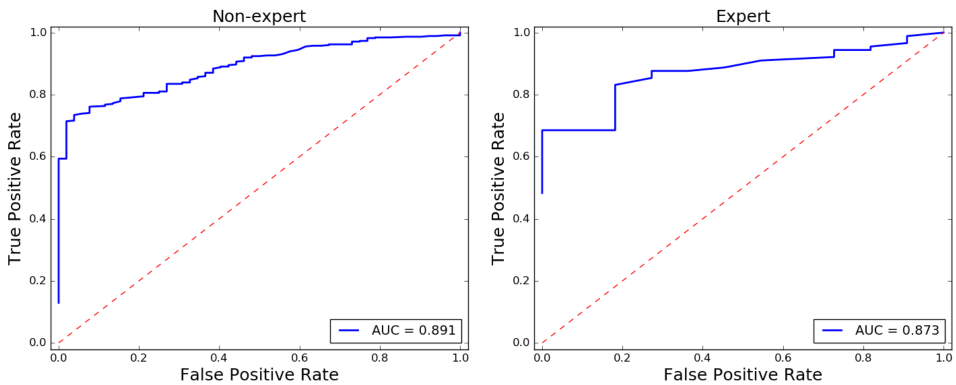
Supplementary Figure 2. The number of questions answered by each expert worker ranged from several to several hundred.

Image Type \ Expertise	Expert	Non-expert
Illumina son	0.80	0.89
Illumina father	0.77	0.88
Illumina mother	0.82	0.9
10X son	0.76	0.77
10X son haplotype 1	0.75	0.70
10X son haplotype 2	0.82	0.69

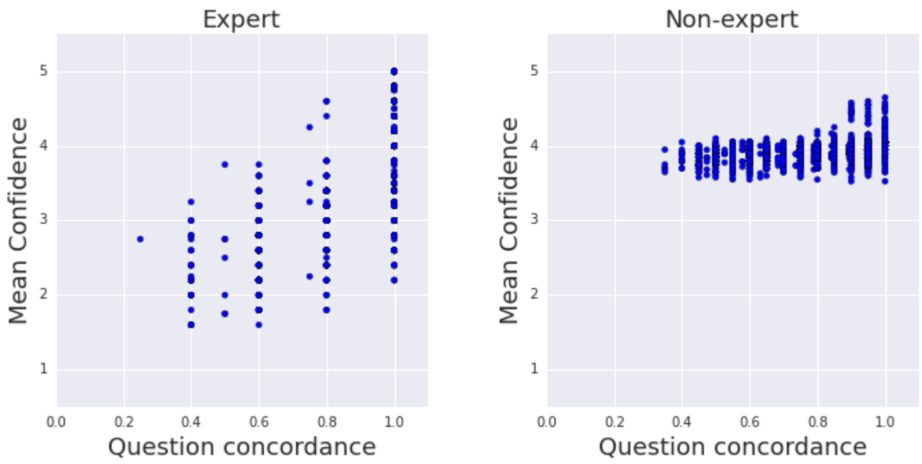
Supplementary Table 2. Fraction of sites with a concordant answer among workers for each image type. A concordant answer was defined as 70% of responses in agreement. We computed concordance rates above this threshold for every image type separately for both expert and non-expert classifications.



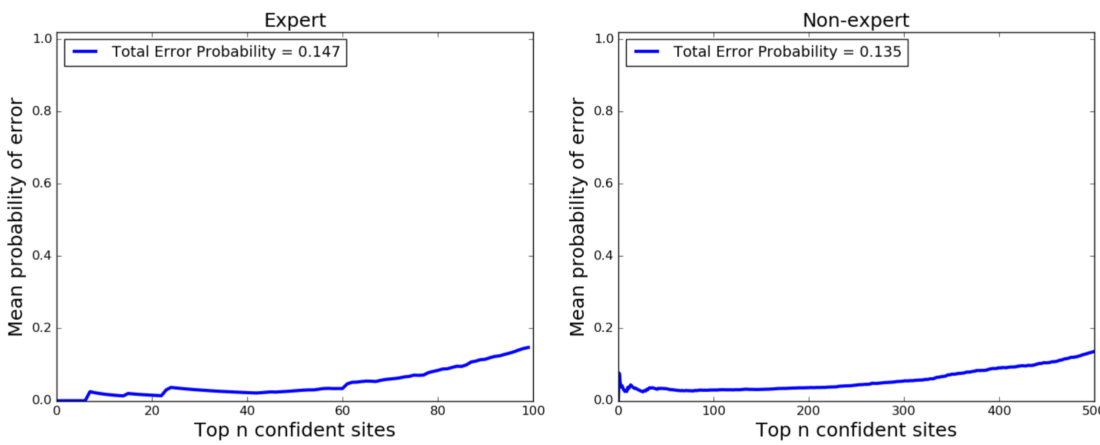
Supplementary Figure 3. Kernel density of percent agreement on most common classification for each image type. Experts and non-experts show similar levels agreement.



Supplementary Figure 4. Non-expert (left) and expert (right) ROC curves discriminating Mendelian violations (1=no violation, 0=violation) sites by the CrowdVariant score.



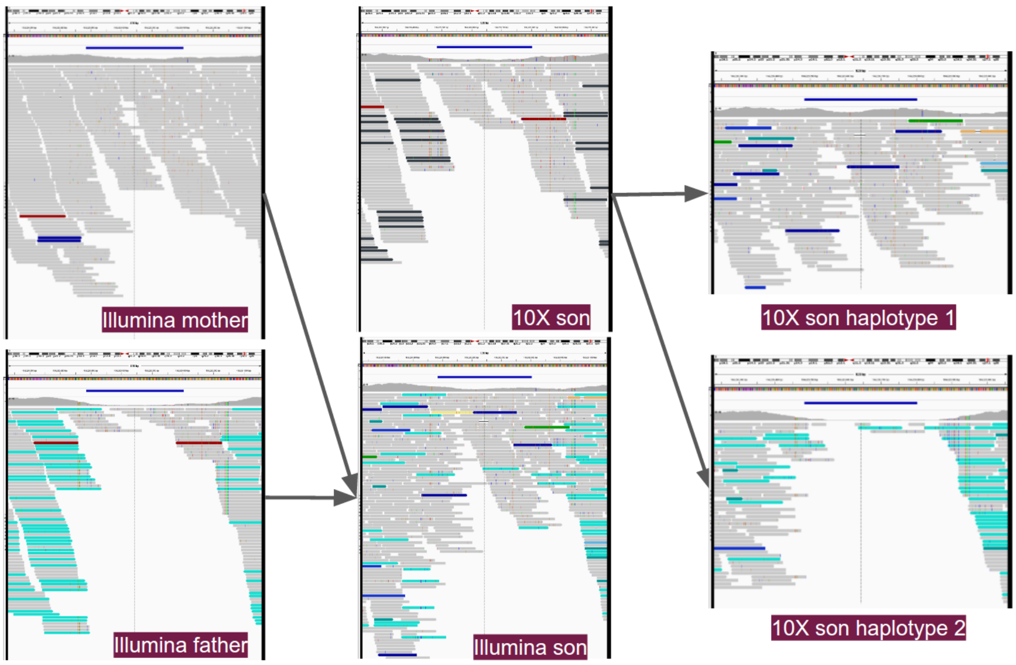
Supplementary Figure 5. Mean reported confidence scores for each question as a function of concordance (percent agreeing with most common answer) for experts [left] and non-experts [right]. Non-expert reported confidence does not display enough variation to sufficiently filter answers based on confidence alone.



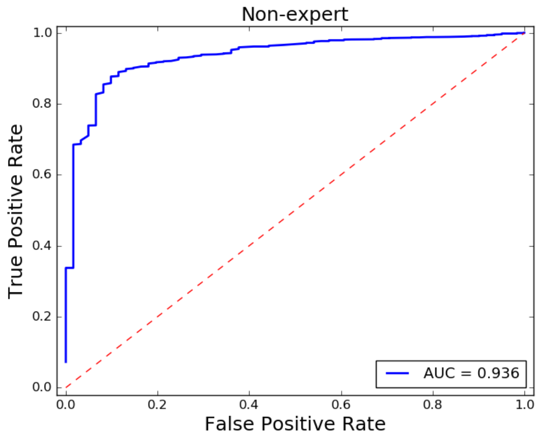
Supplementary Figure 6. Cumulative mean probability of a Mendelian violation among expert [left] and non-expert [right] classifications as a function of the top n sites. We order sites from highest to lowest score [left to right, x-axis] and for the top n, from 1 to the total number of sites, we compute the average probability of an error for those n sites together. For most confident sites the probability is low and increases as confidence decreases.

Metric \ Data Set	Expert Randomized	Non-expert Randomized
Percent of sites without violation	54/100 (54%)	363/500 (73%)
ROC AUC	0.50	0.48
ROC AUC 95% confidence interval	[0.38, 0.61]	[0.42, 0.53]
Average violation probability	0.43	0.31

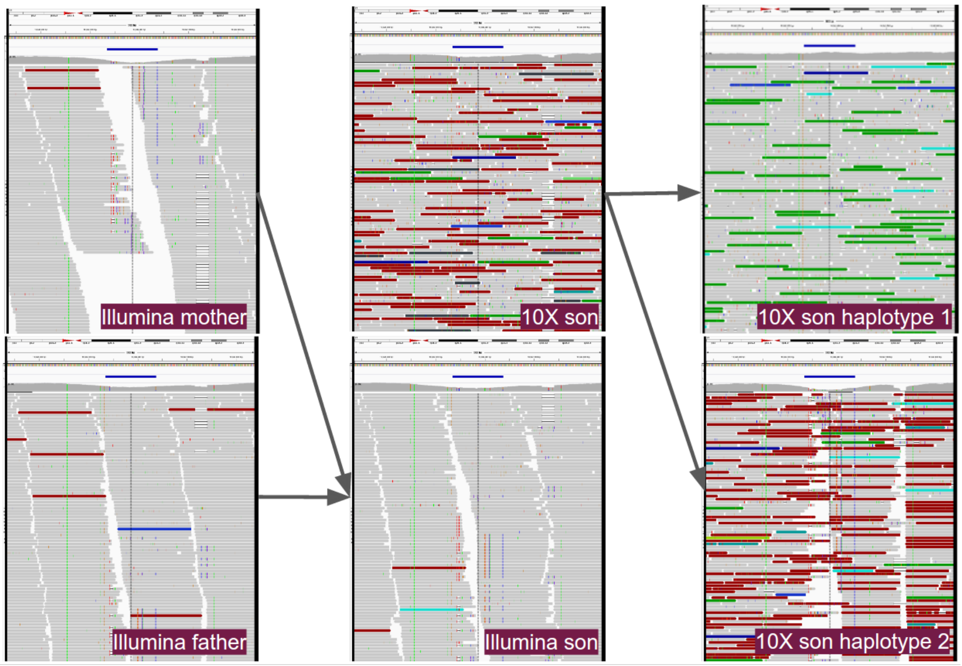
Supplementary Table 3. Randomized answers give a lower bound on violation metrics. We randomized the answer by worker matrix and re-computed evaluation metrics. Both AUC confidence intervals overlap a random AUC of 0.5.



Supplementary Figure 7. Viewing all image types together shows the power of combining familial and phasing information in different sequencing platforms. This variant (chr3:184220484-184220803) was classified as CN1 in the son with CrowdVariant score 0.93 and is part of the gold set. Svviz classified this example as CN2. Clockwise from top left: Illumina mother, 10X son, 10X son haplotype 1, 10X son haplotype 2, Illumina son, Illumina father. Svviz classification: CN2. Crowd classification: CN1.



Supplementary Figure 8. Non-expert genome-wide ROC curves discriminating Mendelian violations (1=no violation, 0=violation) sites by the CrowdVariant score.



Supplementary Figure 9. This variant (chr21:10842335-10842437) was the only Mendelian violation in the genome-wide curated set. It was classified as CN2 for both son and father and CN0 for mother. The CrowdVariant score was 0.94.